

Estimation of Partially Linear Regression Model under Partial Consistency Property

Xia Cui,^{*} Ying Lu[†] and Heng Peng[‡]

Abstract

In this paper, utilizing recent theoretical results in high dimensional statistical modeling, we propose a model-free yet computationally simple approach to estimate the partially linear model $Y = X\beta + g(Z) + \varepsilon$. Motivated by the partial consistency phenomena, we propose to model $g(Z)$ via incidental parameters. Based on partitioning the support of Z , a simple local average is used to estimate the response surface. The proposed method seeks to strike a balance between computation burden and efficiency of the estimators while minimizing model bias. Computationally this approach only involves least squares. We show that given the inconsistent estimator of $g(Z)$, a root n consistent estimator of parametric component β of the partially linear model can be obtained with little cost in efficiency. Moreover, conditional on the β estimates, an optimal estimator of $g(Z)$ can then be obtained using classic nonparametric methods. The statistical inference problem regarding β and a two-population nonparametric testing problem regarding $g(Z)$ are considered. Our results show that the behavior of test statistics are satisfactory.

^{*}School of Mathematics and Information Science, Guangzhou University, Guangzhou, China

[†]Center for the Promotion of Research Involving Innovative Statistical Methodology, Steinhardt

School of Culture, Education and Human Development, New York University, New York, USA

[‡]Department of Mathematics, Hong Kong Baptist University, Hong Kong

To assess the performance of our method in comparison with other methods, three simulation studies are conducted and a real dataset about risk factors of birth weights is analyzed.

Key words and phrases: Partially linear model, Partial consistency, high correlation, categorical data, asymptotic normality, nonparametric testing

AMS 2001 subject classification: 62J05, 62G08, 62G20

1. Introduction

In statistics, regression analysis is a family of important techniques that estimate the relationship between a continuous response variable Y and covariates X with dimension p , $Y = f(X) + \epsilon$. Parametric regression models specify the regression function in terms of a small number of parameters. For example, in linear regression, a linear response surface $E(Y) = X\beta$ is assumed and determined by $p \times 1$ vector of β . The parametric methods are easy to estimate and are widely used in statistical practice as parameters β can be naturally interpreted as the “effects of X on Y”. However the requirement of a pre-determined functional form can increase the risk of model misspecification, which leads to invalid estimates. In contrast, nonparametric methods assume no predetermined functional form and $f(X)$ is estimated entirely using the information from the data. Various kernel methods or smoothing techniques have been developed to estimate $f(X)$. In general, these methods use local information about $f(X)$ to blur the influence of noise at each data point. The bandwidth, h , determines the width of the local neighborhood and the kernel function determines the contribution of data points in the neighborhood. The bandwidth h is essential to the nonparametric estimator $\hat{f}(X)$. Smoother estimates of $f(X)$ are produced as h increases and vice versa. As a special case, the local linear model reduces to linear regression when h spans the entire data set with a flat kernel. The choice of h is data driven and can be computationally demanding as the dimension of X increases. Moreover, nonparametric estimation suffers the curse of dimensionality which requires the sample size to increase exponentially with the dimension of X . In addition, most kernel functions are designed for continuous variables and it is not natural to incorporate categorical predictors. Hence a fully nonparametric approach is rarely useful to estimate the regression function with multiple covariates.

The partially linear model, one of the most commonly used semi-parametric

regression models,

$$Y_i = X_i^\top \beta + g(Z_i) + \varepsilon_i, \quad i = 1, \dots, n \quad (1.1)$$

offer an appealing alternative in that it allows both parametric and nonparametric specifications in the regression function. In this model, the covariates are separated into parametric components $X_i = (X_{i1}, \dots, X_{ip})^\top$ and nonparametric components $Z_i = (Z_{i1}, \dots, Z_{iq})^\top$. The parametric part of the model can be interpreted as a linear model, while the nonparametric part frees the model from stringent structural assumptions. As a result, the estimates of β are also less affected by model bias. This model has gained great popularity since it was first introduced by Engle, Granger, Rice and Weiss (1986) and has been widely applied in economics, social and biological sciences. A lot of work have also been devoted to the estimation of the partially linear models. Engle, Granger, Rice and Weiss (1986) and many others study the penalized least squares method for partially linear regression models estimation. Robinson (1988) introduces a profile least squares estimator for β based on the Nadaraya-Watson kernel estimate of the unknown function $g(\cdot)$. Heckman (1986), Rice (1986), Chen (1988) and Speckman (1988) study the consistency properties of the estimate of β under different assumptions. Schick (1996) and Liang and Härdle (1997) extend the root n consistency and asymptotic results for the case of heteroscedasticity. For models with only specification of the first two moments, Severini and Staniswalis (1994) propose a quasi-likelihood estimation method. Härdle, Mammen and Müller (1998) investigate nonparametric testing problem of the unknown function $g(\cdot)$. Among others, Härdle, Liang and Gao (2000) provide a good comprehensive reference of the partially linear model.

Most of the above methods are based on the idea of first taking the conditional expectation give Z_i and then subtracting the conditional expectations in both sides

of (1.1). This way, the function $g(\cdot)$ disappears,

$$Y_i - \mathbb{E}(Y_i|Z_i) = \{X_i - \mathbb{E}(X_i|Z_i)\}^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n. \quad (1.2)$$

If the conditional expectations were known, $\boldsymbol{\beta}$ could be readily estimated via regression techniques. In practice, the quantities $\mathbb{E}(Y|Z)$ and $\mathbb{E}(X|Z)$ are estimated via nonparametric method. The estimation of these conditional expectations is very difficult when the dimension of Z_i is high. Without accurate and stable estimates of those conditional expectations, the estimates of $\boldsymbol{\beta}$ will also be negatively affected. In fact, Robinson (1988), Andrews (1994) and Li (1996) obtain the root- n consistency of the estimator of $\boldsymbol{\beta}$ under an important bandwidth condition with respect to the nonparametric part: $\sqrt{n}\left(h^4 + \frac{1}{nh^q}\right) \rightarrow 0$. Clearly, this condition breaks down when $q > 3$.

To circumvent the curse of dimensionality, $g(Z)$ is often specified in terms of additive structure of one-dimensional nonparametric functions, $\sum_{j=1}^q g_j(Z_j)$. This is the so-called generalized additive model. In theory, if the specified additive structure corresponds to the underlying true model, every $g_j(\cdot)$ can be estimated with desired one-dimensional nonparametric precision, and $\boldsymbol{\beta}$ can be estimated efficiently with optimal convergent rate. But in practice, estimating multiple nonparametric functions is related to complicated bandwidth selection procedures, which increases computation complexity and makes the results unstable. Moreover, when variables $\{Z_j\}$ are highly correlated, the stability and accuracy of such additive structure in partially linear regression model is problematic (see Jiang, Fan and Fan, 2010). Lastly, if the additive structure is misspecified, for example, when there are interactions between the nonparametric predictors Z , the model and the estimation of $\boldsymbol{\beta}$ will be biased.

In this paper, we propose a simple least squares based method to estimate the parametric component of model (1.1) without complicated nonparametric estima-

tion. The basic idea is as follows. Since the value of $g(Z)$ at each point is only related to the local properties of $g(\cdot)$, it can be represented by a set of incidental parameters that are only related to finite local sample points. Inspired by the partial consistency property (Neyman and Scott, 1942; Lancaster (2000); Fan et al. (2005)), we propose to approximate $g(Z)$ using local averages over small partitions of the support of $g(\cdot)$. The parametric parameters β can then be estimated using profile least squares. Following the classic results about the partial consistency property (Fan et al. (2005)), we show that, under moderate conditions this estimator of β has optimal root- n consistency and is almost efficient. Moreover, given a good estimate of β , an improved estimate of the nonparametric component $g(Z)$ can be obtained. Compared to the classic nonparametric approach, this method is not only easy to compute, it also readily incorporates covariates Z when they contain both continuous and categorical variables. We also explore the statistical inference problems regarding the parametric and nonparametric components under the proposed estimating method. Two test statistics are proposed and their limiting distributions are examined.

The rest of the paper is organized as follows. In section 2, followed by a brief technical review of the partial consistency property, we propose a new estimation method of the parametric component for the partially linear regression model. The consistency of the parameter estimates are shown when the nonparametric component consists of univariate, one continuous and one categorical variable or two highly correlated continuous variables. The inference methods of the partially linear regression model are discussed in Section 3. Numerical studies assessing the performance of the proposed method in comparison with existing alternative methods are presented in Section 4. A real data example is analyzed in Section 5. In Section 6 we offer an in-depth discussion about the implications of the proposed method and

further directions. Technical proofs are relegated to the Appendix.

2. Estimating partially linear regression model under partial consistency property

2.1. Review of partial consistent phenomenon

The partial consistency property refers to a phenomenon when a statistical model contains nuisance parameters whose number grows with sample size; although the nuisance parameter themselves cannot be estimated consistently, the rest of the parameters sometimes can be. Neyman and Scott (1942) first studied this phenomenon. Using their terminology, the nuisance parameters are “incidental” since each of them is only related to finite sample points, and the parameters that can be estimated consistently are “structural” because every sample point contains their information. The partial consistency phenomenon appears in mixed effect models, models for longitudinal data, and panel data in econometrics, see Lancaster (2000) etc. In one JASA discussion paper, Fan, Peng and Huang (2005) formally studied the theoretical properties of parameter estimators under partial consistency and their applications to microarray normalization. They consider a very general form of regression model

$$\mathbf{Y}_n = \mathbf{B}_n \boldsymbol{\alpha}_n + \mathbf{Z}_n \boldsymbol{\beta} + \mathbf{M} + \boldsymbol{\epsilon}_n, \quad n = J \times I, \quad (2.1)$$

where $\mathbf{Y}_n = (Y_1, \dots, Y_n)^T$, $\mathbf{B}_n = \mathbf{I}_J \otimes \mathbf{1}_I$ is an $n \times J$ design matrix, I is assumed to be a constant and J grows with sample size n . \mathbf{Z}_n is an $n \times d$ random matrix with d being the dimension of $\boldsymbol{\beta}$, $\mathbf{M} = (m(X_1), \dots, m(X_n))$ is an nonparametric function, and $\boldsymbol{\epsilon}_n = (\epsilon_1, \dots, \epsilon_n)$ is a vector of i.i.d. errors. In the above model, $\boldsymbol{\alpha}_n$ is a vector of incidental parameters as its dimension J increases with sample size, $\boldsymbol{\beta}$ and \mathbf{M} are

the structural parameters. Fan *et al.* (2005) show that β and M can be estimated consistently and even nearly efficient when the value of I is moderately large,

$$\sqrt{n}(\hat{\beta} - \beta) \sim \mathcal{N}(0, \frac{I}{I-1} \sigma^2 \Sigma^{-1}),$$

the factor $I/(I-1)$ is the price to pay for estimating the nuisance parameters α_n .

2.2. Estimating partially linear model under partial consistency

First we apply our proposed strategy to estimate a partially linear regression model with one-dimensional nonparametric component,

$$Y_i = X_i \beta + g(Z_i) + \varepsilon_i, \quad i = 1, \dots, n, \quad (2.2)$$

where $g(\cdot)$ is an unknown function, $Z_i \in R^1$ is a continuous random variable, and other assumptions for the model are similar as those imposed on the model (2.1). Without loss of generality, we assume that Z_i are i.i.d random variables and follow $[0, 1]$ uniform distribution, and is sorted as $0 \leq Z_1 \leq Z_2 \leq \dots \leq Z_n \leq 1$ based on their realized values. Then we can partition the support of Z_i into $J = n/I$ sub-intervals such that the j th interval covers I different random variables with closely realized values from $z_{(j-1)I+1}$ to z_{jI} . If the density of Z_i is smooth enough, these sub-intervals should be narrow and the values of $g(\cdot)$ over the same sub-interval should be close and $g(Z_{(j-1)I+1}) \approx g(Z_{(j-1)I+2}) \dots \approx g(Z_{jI}) \approx \alpha_j$ where $\alpha_j = \frac{1}{I} \sum_{i=1}^I g(Z_{(j-1)I+i})$. Then the nonparametric part of model (2.2) can be reformulated in terms of partially consistent observations and rewritten in the form of the model (2.1)

$$\mathbf{Y}_n = \mathbf{B}_n \alpha_n + \mathbf{X}_n \beta + \varepsilon_n^*, \quad n = J \times I \quad (2.3)$$

with $\varepsilon_{(j-1)I+i}^* = \varepsilon_{(j-1)I+i} + g(Z_{(j-1)I+i}) - \frac{1}{I} \sum_{i=1}^I g(Z_{(j-1)I+i})$. It is easy to see that the second term in $\varepsilon_{(j-1)I+i}^*$ is the approximation error. Normally when I is a small

constant, it is of order $O(1/J)$ or $O(1/n)$, and much smaller than ε . Hence the approximation error can be ignored and it is expected that, similar to as in (2.1), β in the model (2.2) or (2.3) can be estimated almost efficiently even when $g(\cdot)$ in (2.2) is not estimated consistently.

Model (2.3) can be easily estimated by profile least squares,

$$\sum_{j=1}^J \sum_{i=1}^I (Y_{(j-1)I+i} - X_{(j-1)I+i} \beta - \alpha_j)^2. \quad (2.4)$$

the estimates of β and α_j can be expressed as follows,

$$\begin{cases} \hat{\beta} = \left\{ \sum_{j=1}^J \sum_{i=1}^I \left\{ X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i} \right\}^T \left\{ X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i} \right\} \right\}^{-1} \\ \quad \times \left\{ \sum_{j=1}^J \sum_{i=1}^I \left\{ X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i} \right\}^T \left\{ Y_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I Y_{(j-1)I+i} \right\} \right\}, \\ \hat{\alpha}_j = \frac{1}{I} \sum_{i=1}^I \left\{ Y_{(j-1)I+i} - X_{(j-1)I+i} \hat{\beta} \right\}. \end{cases} \quad (2.5)$$

We have the following theorem for the above profile least squares estimate of β under the model (2.2) or (2.3).

Theorem 1. *Under regularity conditions (a)–(d) in the Appendix, for the profile least squares estimator of β defined in (2.5),*

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} N(0, \frac{I}{I-1} \sigma^2 \Sigma^{-1}), \quad (2.6)$$

where $\Sigma = \mathbb{E} \left[\{X - \mathbb{E}(X|Z)\} \{X - \mathbb{E}(X|Z)\}^\top \right]$.

Similar to the treatment of least square estimator for linear regression models, and noting that the degrees of freedom of (2.3) is approximately $(I-1)/I \cdot n$, we can estimate the variance of $\hat{\beta}$ using sandwich formula based on (2.5).

$$\text{Var}(\hat{\beta}) = \hat{\sigma}^2 \left\{ \sum_{j=1}^J \sum_{i=1}^I \left\{ X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i} \right\}^T \left\{ X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i} \right\} \right\}^{-1}, \quad (2.7)$$

where

$$\hat{\sigma}^2 = \frac{I}{I-1} \cdot \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^I (Y_{(j-1)I+i} - X_{(j-1)I+i} \hat{\beta} - \hat{\alpha}_j)^2.$$

Furthermore, we can plug $\hat{\beta}$ back into equation (2.2) and obtain an updated nonparametric estimate of $g(Z)$ based on

$$Y_i^* = Y_i - X_i \hat{\beta}$$

using standard nonparametric techniques. Since $\hat{\beta}$ is a root n consistent estimator of β , we expect the updated nonparametric estimator $\hat{g}(Z)$ will converge to $g(Z)$ at the optimal nonparametric convergence rate.

2.3. Extension to multivariate nonparametric $g(Z)$

Case I: The simple method of approximating one-dimensional function $g(Z_i)$ can be readily extended to the multivariate case when Z consists of one continuous variable and several categorical variables. Note that without loss of generality, we can express multiple categorical variables as one K -level categorical variable. Hence, a partially linear model

$$Y_i = X_i \beta + g(Z_i^d, Z_i^c) + \varepsilon_i, \quad i = 1, \dots, n, \quad (2.8)$$

where $Z_i = (Z_i^d, Z_i^c)$ where $Z_i^c \in R^1$ as specified in (2.2), Z_i^d is a K -level categorical variable.

To approximate $g(Z^d, Z^c)$ we first split the data into K subsets given the categorical values of Z_i^d , then the k th ($0 \leq k \leq K$) subset of the data will be further partitioned into sub-intervals of I data points with adjacent values of Z^c . Based on the partition, model (2.8) can still be written in the form of (2.3). The profile least squares as shown above can be used to estimate β and we have the following corollary.

Corollary 1. *Under the model (2.8) and regularity conditions (a)–(e), for the profile least squares estimator of β defined in (2.5),*

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} N(0, \frac{I}{I-1} \sigma^2 \Sigma^{-1}), \quad (2.9)$$

where $\Sigma = \mathbb{E} \left[\{X - \mathbb{E}(X|Z)\} \{X - \mathbb{E}(X|Z)\}^\top \right]$.

Case II: The simple approximation can also be easily applied to continuous bivariate variable $Z = (Z_1, Z_2) \in \mathbb{R}^2$. The partition will need to be done over the bivariate support of Z . In the extreme case when the two components of $Z = (Z_1, Z_2)$ are independent from each other, the approximation error based on the partition is of order $o(1/\sqrt{n})$, the same as the model error. Hence in theory the root- n consistency of β can be established. Below we outline a corollary that based on the case when the two components of Z are highly correlated so we only need to partition the support of Z according to one component. First we assume

$$\Delta_{si} \equiv Z_{1i} - Z_{2i} \rightarrow 0, \quad i = 1, \dots, n, \quad (2.10)$$

a similar condition as in Jiang, Fan and Fan (2010)

Under the assumption (2.10) with $\Delta_{si} = o(1)$, it is sufficient to partition the observations into subintervals of I data points according to the order of $Z_{1i}, i = 1, \dots, n$. If $g(\cdot)$ satisfies some regular smoothness conditions, given subinterval j , $g(\mathbf{Z}_{(j-1)I+i})$ is approximately equal for $i = 1, \dots, I$, denoted by α_j . Again the model can be represented in the form of (2.3) and we have another corollary,

Corollary 2. *Under the model (2.8) where Z_{1i} and Z_{2i} are highly correlated and satisfy the condition (2.10), and the regularity conditions (a)–(d), for the profile least squares estimator of β defined in (2.5),*

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} N(0, \frac{I}{I-1} \sigma^2 \Sigma^{-1}), \quad (2.11)$$

where $\Sigma = \mathbb{E} \left[\{X - \mathbb{E}(X|Z)\} \{X - \mathbb{E}(X|Z)\}^\top \right]$.

The results of the theorem and corollaries are similar as the results of Fan *et al.* (2005) except replacing the unconditional asymptotic covariance matrix of the estimate by the conditional covariance matrix. The proofs of the theorem and corollaries are deferred to the Appendix.

Remark 1: We proposed to approximate $g(Z)$ by simply averaging observations within the local neighborhood. This method to some extent resembles kernel methods with small bandwidth in nonparametric estimation. However, our method does not require a kernel density function nor complicated bandwidth selection, so it can be viewed as a “poor man’s” nonparametric method that is completely model free. Theorem 1 and the two corollaries demonstrate that the limiting distribution of $\sqrt{n}(\hat{\beta} - \beta)$ based on partially consistent estimation of $g(Z)$ is almost as efficient as the estimator of β based on classic method for partially linear model while the latter requires a consistent estimates of $g(Z)$. Theorem 1 shows that the parametric estimates based on simply a naive approximation of $g(Z)$ can still obtain optimal root- n consistency. In the extreme case, the consistent estimate of β can be obtained when the number of observations per subinterval, I , is as small as 2. One only pays the cost in efficiency by a factor of $I/(I - 1)$. This inflation factor diminishes quickly as I increases.

Remark 2: Computationally, our method is easy to compute and does not require additional tuning parameter selection. In the simulation section we will show the complex computational procedure of the nonparametric kernel method also leads to numerical inefficiency, the ratio between the average estimation errors of our method and the nonparametric kernel method is in fact less than $I/(I - 1)$. In addition, the effectiveness of various kernel functions only depends on the underlying assumption about $g(Z)$. In our method, $g(Z)$ is approximated by non-overlapping partitions hence it is more local than the kernel methods, and therefore it is more forgiving to

oddities such as jumps, singularities and boundary effects of $g(Z)$. Essentially, we do not cast any structural assumption over $g(Z)$ so it can be more readily extended to deal with multivariate random vectors including ordered and unordered categorical data and allow interactions among the components.

Remark 3: As the dimension of the continuous components of Z increases, similar as the discussion of Fan and Huang (2001) about ordering multivariate vector, Z can be ordered according to the first principle component of Z or certain covariate. In practice, as shown by Cheng and Wu (2013), the high dimensional continuous random vector Z can often be represented by a low dimensional manifold. Hence we can expect that for many cases, once Z is expressed in a low dimensional manifold without losing much information, the partition of Z can be done within the manifold effectively and our results should still apply. Nevertheless, further investigation is needed to ascertain the conditions needed for the generalization of our method.

3. Statistical inference for partially linear regression model

3.1. Statistical inference for parametric component

In this section, we investigate statistical inference problem with respect to the estimator of β . In particular, we consider the following testing problem for β

$$H_0^1 : A\beta = 0, \quad \text{vs} \quad H_1^1 : A\beta \neq 0 \quad (3.12)$$

where A is a $k \times p$ matrix. A profile likelihood ratio or profile least square ratio test statistic (Fan and Huang, 2005) will be defined and we will investigate whether this test statistic is almost efficient and has an easy-to-work limiting distribution

Let $\hat{\beta}_0$ be the estimators of β and $\hat{\alpha}_{n0}$ be the estimators of α_n in (2.3) under the null hypothesis H_0^1 . The residual sum of squares (RSS) under the null hypothesis is $RSS_0 = n^{-1} \sum_{j=1}^J \sum_{i=1}^I \hat{\varepsilon}_{(j-1)I+i,0}^2$, where $\hat{\varepsilon}_{(j-1)I+i,0} = Y_{(j-1)I+i} - \hat{\alpha}_{j0} - X_{(j-1)I+i} \hat{\beta}_0$. Similarly, let $\hat{\beta}_1$ and $\hat{\alpha}_{n1}$ be the estimators of β and α_n in (2.3) under the alternative hypothesis. The RSS under H_1^1 is $RSS_1 = n^{-1} \sum_{j=1}^J \sum_{i=1}^I \hat{\varepsilon}_{(j-1)I+i,1}^2$, where $\hat{\varepsilon}_{(j-1)I+i,1} = Y_{(j-1)I+i} - \hat{\alpha}_{j1} - X_{(j-1)I+i} \hat{\beta}_1$. Following Fan and Huang (2005), we define a profile least squares based test statistic

$$T_n^1 = (RSS_0 - RSS_1) / RSS_1. \quad (3.13)$$

Then under the regularity conditions and the null hypothesis, we have the following theorem for the asymptotic distribution of T_n^1 .

Theorem 2. *Under regularity conditions (a)–(e) in the Appendix, and given the profile least squares estimator $\hat{\beta}_0$ and $\hat{\beta}_1$ defined above,*

$$\frac{I-1}{I} \cdot nT_n^1 \xrightarrow{\mathcal{L}} \chi_k^2 \quad \text{as } n \rightarrow \infty. \quad (3.14)$$

For linear regression with normal errors, the same test statistic was shown to have a Chi-square distribution (Fan and Huang, 2005). The results in Theorem 2 demonstrates that classic hypothesis testing results can still be applied to the parametric component in (2.2) with partially consistent nonparametric component estimators. The constant $I/(I-1)$ is the price to be paid for introducing high dimensional nuisance parameters in the model.

In practice, for finite sample size, we can use the following bootstrap procedure to calculate the p -value of the proposed testing statistic T_n^1 under null hypothesis.

Bootstrap algorithm for T_n^1

1. Generate the residuals $\{\hat{\varepsilon}_{(j-1)I+i}^*, j = 1, \dots, J; i = 1, \dots, I\}$ by uniformly resampling from $\{\hat{\varepsilon}_{(j-1)I+i,0}\}$, then centralize $\{\hat{\varepsilon}_{(j-1)I+i}^*\}$ to have mean zero.

2. Define the bootstrap sample $Y_{(j-1)I+i}^* = X_{(j-1)I+i}\hat{\beta}_0 + \hat{\alpha}_{j,0} + \hat{\varepsilon}_{(j-1)I+i}^*$.
3. Calculate the bootstrap test statistics T_n^{1*} based on the bootstrap sample $\left\{ (Y_{(j-1)I+i}^*, X_{(j-1)I+i}, Z_{(j-1)I+i}), j = 1, \dots, J; i = 1, \dots, I \right\}$.
4. Repeat steps 1-3 to obtain N replicates of bootstrap samples and compute $T_n^{1*,b}$ for each sample $b = 1, \dots, N$. The p -value of the test can be calculated based on the relative frequency of the events $\{T_n^{1*,b} \geq T_n^1\}$.

3.2. Statistical inference for nonparametric component when categorical data are involved

In Corollary 1, we established the root- n consistency of the parametric component β when the nonparametric component is of the form $Z_i = (Z_i^d, Z_i^c)$ where Z_i^d is a N -level categorical variable and Z_i^c is a continuous variable in R^1 .

Given the almost efficient estimate $\hat{\beta}$, we have $Y_i^* = Y_i - X_i\hat{\beta} = g(Z_i^d, Z_i^c) + \varepsilon_i^*$. The nonparametric function $g(Z_i) = g(Z_i^d, Z_i^c)$ can be expressed in terms of a series of univariate functions conditioning on the values of Z_i^d , $g(Z_i^c | Z_i^d = k)$, $k = 1, \dots, N$. Each of these univariate functions can be estimated using kernel method based on the split data with corresponding Z_i^c values. Those estimates can be defined as $\hat{g}(Z_i^c | Z_i^d = k)$. Naturally one likes to test the equivalence of these univariate functions.

Motivated by the real example in Section 5, we consider the following testing problem when $N = 2$:

$$\begin{aligned} H_0^2 : g(Z_i^c | Z_i^d = 0) &= g(Z_i^c | Z_i^d = 1) \text{ almost everywhere,} \\ H_1^2 : g(Z_i^c | Z_i^d = 0) &\neq g(Z_i^c | Z_i^d = 1) \text{ on a set with positive measure.} \end{aligned} \tag{3.15}$$

The above testing problem resembles a two-population nonparametric testing problem. For such a testing problem, Racine, Hart and Li (2006) suggest a quadratic

distance testing statistic. However, the quadratic distance statistics are not sensitive to the local changes. Based on L_∞ norm and the idea from Fan and Zhang (2000), we suggest the following statistic in the context of partially linear model.

$$T_n^2 = (-2 \log h)^{1/2} \left[\sup_{Z^c} \frac{|\hat{g}(Z^c|Z^d = 1) - \hat{g}(Z^c|Z^d = 0)|}{\sqrt{\widehat{\text{Var}}\{\hat{g}(Z^c|Z^d = 1) - \hat{g}(Z^c|Z^d = 0)\}}} - d_n \right], \quad (3.16)$$

where h is the chosen bandwidth parameter when estimating $g(Z^c|Z^d)$ and

$$d_n = (-2 \log h)^{1/2} + \frac{1}{(-2 \log h)^{1/2}} \log \left\{ \frac{\int K^2(t) dt}{4\pi \int K^2(t) dt} \right\},$$

with $K(\cdot)$ is a kernel function satisfying $\int K(t) dt = 1$ and $\int t^2 K(t) dt > 0$.

Notice that $\hat{g}(Z^c|Z^d = 1)$ and $\hat{g}(Z^c|Z^d = 0)$ are estimated by different samples, hence $\hat{g}(Z^c|Z^d = 1)$ and $\hat{g}(Z^c|Z^d = 0)$ can be assumed independent. So

$$\text{Var}\{\hat{g}(Z^c|Z^d = 1) - \hat{g}(Z^c|Z^d = 0)\} = \text{Var}\{\hat{g}(Z^c|Z^d = 0)\} + \text{Var}\{\hat{g}(Z^c|Z^d = 1)\},$$

where $\text{Var}\{\hat{g}(Z^c|Z^d = 0)\}$ and $\text{Var}\{\hat{g}(Z^c|Z^d = 1)\}$, which can be estimated using standard nonparametric procedures.

Given the level of the test, when T_n^2 is greater than the critical value, H_0^2 can be rejected. In general, critical values can be determined by the asymptotical distribution of test statistic under the null hypothesis. However, for this kind of non-parametric testing problem the test statistic tends to converge to its asymptotic distribution very slowly (Racine et al. (2006).) The best way to approximate the null hypothesis distribution for the above testing statistic is by bootstrapping. Following the idea of Racine, Hart and Li (2006), we suggest a simple bootstrap procedure to approximate the null hypothesis distribution of T_n^2 .

Bootstrap algorithm for T_n^2 :

1. Randomly select Z_i^{d*} from $\{Z_i^d, i = 1, \dots, n\}$ with replacement, and call $\{Y_i, X_i, Z_i^{d*}, Z_{i2}\}$ the bootstrap sample.

2. Use the bootstrap sample to compute the bootstrap statistic T_n^{2*} , which is the same as T_n^2 except that Z_{i1} is replaced by Z_{i1}^* values.
3. Repeat steps 1 and 2 to obtain N replicates of bootstrap samples and $T_n^{2*,b}, b = 1, \dots, N$. The p -values is based on the relative frequency of the event $\{T_n^{2*,b} \geq T_n^2\}$ in the replications of the bootstrap sampling.

The distribution of T_n^2 under H_0^2 is asymptotically approximated by the bootstrap distribution of T_n^{2*} . Now let $Q_{1-\alpha}(T_n^{2*})$ be the $(1-\alpha)$ th quantile of the bootstrapped test statistic distribution, the empirical $(1-\alpha)$ confidence band for $\{\hat{g}(Z^c|Z^d = 1) - \hat{g}(Z^c|Z^d = 0)\}$ can be constructed as follows,

$$\left[\{\hat{g}(Z^c|Z^d = 1) - \hat{g}(Z^c|Z^d = 0)\} - \Delta_\alpha(Z^c), \hat{g}(Z^c|Z^d = 1) - \hat{g}(Z^c|Z^d = 0) + \Delta_\alpha(Z^c) \right], \quad (3.17)$$

where

$$\Delta_\alpha(Z_2) = \{d_n + Q_{1-\alpha}(T_n^{2*})(-2 \log h)^{-1/2}\} \sqrt{\widehat{\text{Var}}\{\hat{g}(Z^c|Z^d = 1) - \hat{g}(Z^c|Z^d = 0)\}}.$$

4. Numerical studies

We conduct three simulation examples to examine the effectiveness of the proposed estimation method and testing procedures for the partially linear regression model. The first example is a simple partial linear regression model with one dimensional nonparametric component. In the second example we consider highly correlated bivariate nonparametric components, while in the third one, the nonparametric components are mixed with one categorical and one continuous variable.

To assess estimation accuracy of the parametric components, we compute mean square error, $\text{MSE}(\hat{\beta}) = \sum_{l=1}^p (\hat{\beta}_l - \beta_l)^2$, and the average estimation errors, $\text{ASE}(\hat{\beta}) = \sum_{l=1}^p |\hat{\beta}_l - \beta_l|$. The robust standard deviation estimate (RSD) of $\hat{\beta}$ is calculated using

$(Q_3 - Q_1)/1.349$ where Q_1 and Q_3 are the 25% and 75% percentiles, respectively. The limiting distributions of the test statistics T_n^1 and T_n^2 under the null hypothesis will be simulated. The power curve of each test will be constructed as well. Varying the sample size and the size of the subintervals, I , the performance of our proposed estimation and inference methods will be examined and compared with alternative methods.

For comparison purposes, all the simulations examples are also calculated using available R packages. Package “gam” is used to fit generalized additive model, package “NP” is used to fit nonparametric regression and package “locfit” is used for nonparametric curve fitting. Generalized cross validation method is used to select the optimal bandwidth whenever it is applicable.

Example 1. *Consider the following simple partially linear regression model*

$$Y_i = X_i^\top \beta + g(Z_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

where $\beta = (1, 3, 0, 0, 0, 0)$ and $g(Z_i) = 3 \sin(2Z_i)$. $(X_i, Z_i), i = 1, \dots, n$ are i.i.d. draws from a multivariate normal distribution with mean zero and the covariance matrix

$$\begin{pmatrix} 1.0 & \rho & \cdots & \rho \\ \rho & 1.0 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1.0 \end{pmatrix}.$$

with $\rho = 0.5$. $\varepsilon_i, i = 1, \dots, n$ are i.i.d. and follow the standard normal distribution.

For this example, 400 simulated samples are produced to evaluate the performance of the proposed parameter estimators. The results will be compared with those produced by function `gam` in R package “gam” that fits Generalized Additive Models.

First the theoretical results of Theorem 1 are nicely illustrated in the left graph of Figure 1 by a linear relationship between $\log(MSE) - \log(I/(I-1))$ and logarithm of the sample size with a slope close to -1. Moreover, from Table 1, when the size of the subintervals is moderately large (e.g. $I \geq 5$), the average estimation errors (ASE) and the estimated standard error of our proposed method are comparable with the results based on `gam` function in R. When sample size increases, the proposed method are better. In the extreme case when $I = 2$, the ASE decreases with sample size and it is only about 1.3 times that of function `gam`. This implies the empirical model variance of our method is about 1.7 times that of `gam` results. This and similar results in the other two numerical studies suggest that although in theory our method has an efficiency loss by a factor of $I/(I-1)$, in practice the kernel based methods also suffer efficiency loss due to computational complexity that is not captured in theoretical results.

We also carry out T_n^1 to test the following null hypothesis:

$$H_0^1 : \beta_3 = \beta_4 = \cdots = \beta_p = 0.$$

We examine the size and power of T_n^1 by producing random samples from a sequence of alternative hypothesis models indexed by parameter δ_1 as follows:

$$H_1^1 : \beta_3 = \delta_1, \quad \beta_l = 0 \quad \text{for } l \geq 4.$$

δ_1 takes values from the set $(0, 1)$. When $\delta_1 = 0$, the alternative hypothesis becomes the null hypothesis. The empirical null distribution of $I/(I-1)nT_n^1$ with $I = 2$ for different sample sizes are calculated based on 1000 simulated samples and plotted in the right panel of Figure 1. We can see that, even for small value of $I = 2$, the empirical null distribution gets closer to the asymptotical distribution χ_4^2 (solid line) as sample size increases. This is consistent with Theorem 2.

To assess the bootstrap procedures proposed in section 3.1, we generate 1000 bootstrap samples and calculate the p -value of the test for each simulated sample. Figure 2 illustrates the behavior of the power functions with respect to different δ values and I values. Two sample sizes are considered, the left panel $n = 100$, and the right panel $n = 200$. Though small value of I increases the variance of the estimator, the power of the test T_n^1 is not compromised. As shown in Figure 2, the power curves are similar for different values of I . The simulation results further confirm that the profile least squares test statistic T_n^1 is a useful tool for linear testing problem in the partially linear regression model under partial consistency

Table 1: Average Estimation Errors for Simulation Example 1 (estimated standard errors in parentheses)

Method	Our Method				GAM
	I=2	I=5	I=10	I=20	
n=100	0.969(0.316)	0.745(0.258)	0.856(0.284)	1.095(0.371)	0.723 (0.231)
n=200	0.662(0.227)	0.520(0.185)	0.479(0.156)	0.579(0.177)	0.501(0.153)
n=400	0.456(0.137)	0.344(0.108)	0.333(0.118)	0.347(0.123)	0.348(0.121)

Example 2. Consider the following generalized additive model,

$$Y_i = X_i^\top \beta + g_1(Z_{1i}) + g_2(Z_{2i}) + g_3(Z_{3i}) + \varepsilon_i, \quad i = 1, \dots, n$$

where parameter β equals to $(1.5, 0.3, 0, 0, 0, 0)^\top$. The functions g_1, g_2, g_3 are:

$$g_1(Z_{1i}) = -5 \sin(2Z_{1i}), \quad g_2(Z_{2i}) = (Z_{2i})^2 - 2/3, \quad g_3(Z_{3i}) = Z_{3i}.$$

X_i follows a multivariate normal distribution with mean vector zero and the covari-

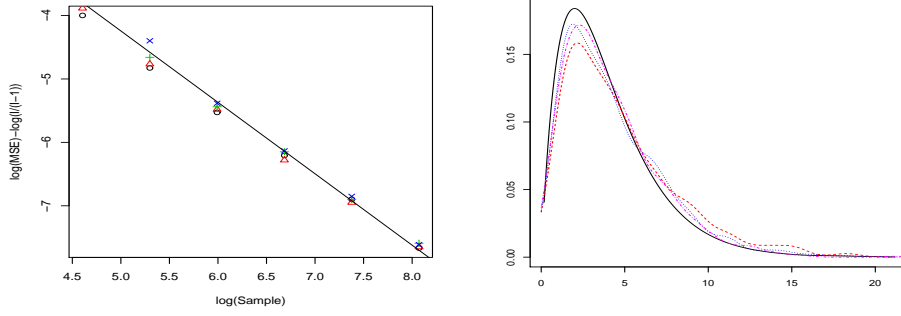


Figure 1: Example 1, Left: Plots of MSEs of β : $I=2$ (\circ), $I=5$ (Δ), $I=10$ ($+$), $I=20$ (\times). The slope of the regression line between $\log(\text{MSE}) - \log(I/(I-1))$ and $\log(\text{Sample})$ is -1.12615. Right: Estimated density of the scaled test Statistic $I/(I-1)nT_n^1$ for $n = 100$ (long-dash), $n = 200$ (dot) and $n = 400$ (dot-dash) with the χ_4^2 distribution (solid) when $I = 2$.

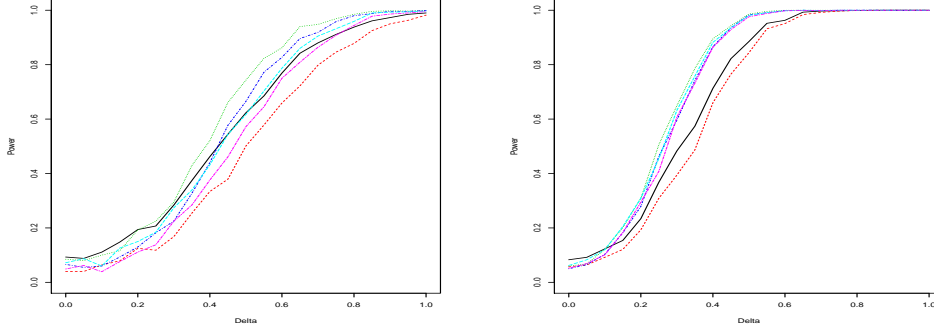


Figure 2: Power of T_n^1 for Example 1, Left: $n = 100$, Right $n = 200$. Solid line($I=2$), Dot line ($I=5$) and Long dash line ($I=10$) are power curves based on scaled $\chi^2(4)$ distribution. Short dash line ($I=2$), Dot-Short dash line ($I=5$), Dot-long dash line ($I=10$) are power curves based on the bootstrap algorithm for T_n^1 .

ance matrix as in Example 1. The Z are constructed to be highly correlated.

$$Z_1 = X_1 + N(0, 1)$$

$$Z_2 = Z_1 + n^{-1/2}u_1$$

$$Z_3 = Z_1 + n^{-1/2}u_2$$

where n is the sample size and u_s ($s = 1, 2$) are $N(0, 1)$ variables independent of

the covariates. The correlation of Z therefore goes up with sample size. Finally the error term $\varepsilon_i \sim N(0, 1)$.

Table 2: The Average Estimation Errors for Example 2 (estimated standard errors in parentheses)

Method	Our Method				GAM
	I=2	I=5	I=10	I=20	
n=100	1.375(0.569)	1.305(0.482)	1.636(0.580)	2.463(0.865)	1.134 (0.454)
n=200	0.741(261)	0.738(0.257)	0.876(0.317)	1.093(0.401)	0.791(0.291)
n=400	0.519(0.184)	0.432(0.161)	0.473(0.165)	0.596(0.221)	0.562(0.213)

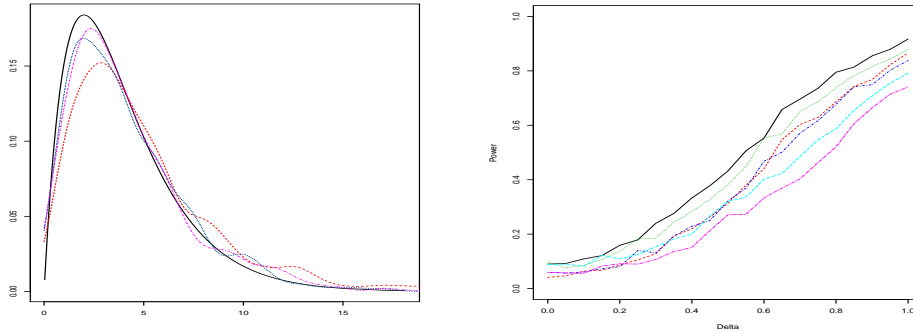


Figure 3: Example 2, Left: Estimated density of the Test Statistics $I/(I-1)nT_n^1$ for $n = 100$ (long-dash) , $n = 200$ (dot) and $n = 400$ (dot-dash) with the χ_4^2 distribution (solid) when $I = 2$. Right: Power of T_n^1 , Solid line(I=2), Dot line (I=5) and Long dash line (I=10) are power curves based on χ_4^2 distribution. Short dash line (I=2), Dot-Short dash line (I=5), Dot-long dash line (I=10) are power curves based on the bootstrap algorithm for T_n^1 .

As in Example 1, 400 simulation examples are used to evaluate the performance of the proposed estimating method. One thousand simulation examples and the same number of bootstrap samples are used to study the properties of T_n^1 for the

same testing problem investigated in Example 1. As indicated by Table 2, as sample size increases, our proposed method outperforms the **gam** package even when $I = 2$. In general, we can see that the proposed method is not sensitive to the choice of I as long as it is not chosen to be too large a value relative to the sample size. Given a fixed sample size, larger I will yield smaller number of subintervals and lead to coarser approximation of the nonparametric function. The empirical null distribution of $(I - 1)/InT_n^1$ in comparison with χ_4^2 is shown in Figure 3. It can be seen that as in Example 1, the empirical null distribution is a reasonable approximation of the asymptotical null distribution χ_4^2 . This is true for various values of I . It is also consistent with the result of Theorem 2.

Compared with the results in Example 1, additional nonparametric component increases the estimation variability for our proposed method and the method of GAM. ASE and standard errors are larger in Example 2. It also reduces the power of T_n^1 for the same testing problem as shown in the right graph of Figure 2. However, our proposed method is more robust to the high correlation situation as it is able to produce more efficient results than **gam** when sample size increases.

Example 3. *The model is*

$$Y_i = X_i^\top \beta + g(Z_i^d, Z_i^c) + \varepsilon_i, \quad i = 1, \dots, n.$$

where

$$g(Z_i^d, Z_i^c) = (Z_i^c)^2 + 2Z_i^c + 0.25Z_i^d e^{-16Z_i^{c2}}.$$

and the true parameter β is a 6×1 vector and equals to $(3.5, 1.3, 0, \dots, 0)^\top$. $X_i, i = 1, \dots, n$ are independently generated from Bernoulli distribution with equal probability being 0 or 1. The categorical variable Z_i^d is a Bernoulli variable independent of X_i with $P(Z_i^d = 1) = 0.7$. The variable Z_i^c is continuous and sampled from

a uniform distribution on $[-1, 1]$ and independent of X_i and Z_i^d . The error term $\varepsilon \sim N(0, 0.2^2)$.

For comparison purpose, we use R package `np` to estimate the bivariate function $g(Z_i^d, Z_i^c)$ nonparametrically. In addition, we also use package `gam` to estimate a “pseudo” model with an additive nonparametric structure specified as below,

$$g(Z_i^d, Z_i^c) = \delta Z_i^d + g(Z_i^c) + \varepsilon_i, \quad i = 1, \dots, n.$$

The true nonparametric components are plotted in the left panel of Figure 4. We can see that the “pseudo” model misspecifies the nonparametric components. It will be interesting to compare the performance of the proposed method, generalized additive model and nonparametric method in terms of estimation of the parametric parameter β .

Again, we produced 400 samples for numerical comparison. Table 3 presents the ASE and estimates of β under three different methods. The `np` method tries to estimate β and the bivariate function $g(Z_1, Z_2)$ simultaneously which involves iterative algorithm and complicated tuning parameter selections. Hence we expect the numerical performance will be compromised to some extent. As the other two simulation studies suggested, our method in general produces slightly bigger ASE than the `np` method but in a factor less than $I/(I - 1)$. On the other hand our method produces more precise estimates of β than the nonparametric approach. It is interesting that the GAM approach outperforms the nonparametric approach even under the wrong model specification. In the left panel of Figure 4, we can see that the difference between curves $g(Z_i^c, Z_i^d = 0)$ and $g(Z_i^c, Z_i^d = 1)$ is small relative to the noise hence the more parsimonious specification of the nonparametric part to some extent improves the parametric estimation. However, under the GAM model wrong inference regard the nonparametric components will be made.

In Table 4 we compare the empirical standard deviation of $\hat{\beta}$ (SD_m) with the

one calculated under proposed sandwich formula (2.7)(SD). It is obvious that our proposed formula provide a consistent estimate of the standard deviation of the estimate $\hat{\beta}$.

Table 3: Fitting Results of ASE and Estimation of β for Example 3 based on the proposed method, NP and GAM (estimated standard errors in parentheses)

Method		Our Method				NP	GAM
n		I=2	I=5	I=10	I=20		
100	ASE	0.302 (0.104)	0.298(0.091)	0.367(0.118)	0.505(0.165)	0.254 (0.091)	0.217 (0.078)
	β_1	3.504(0.067)	3.502(0.063)	3.506(0.076)	3.498 (0.112)	3.464 (0.058)	3.499(0.047)
	β_2	1.305 (0.067)	1.294(0.065)	1.302(0.086)	1.310(0.100)	1.291(0.055)	1.302 (0.047)
200	ASE	0.197(0.064)	0.163(0.052)	0.187(0.059)	0.242(0.072)	0.153 (0.052)	0.149(0.048)
	β_1	3.502(0.045)	3.497 (0.033)	3.498(0.042)	3.504 (0.055)	3.486 (0.032)	3.499(0.030)
	β_2	1.300(0.041)	1.299(0.035)	1.303(0.039)	1.297 (0.054)	1.293(0.031)	1.299(0.032)
400	ASE	0.138 (0.041))	0.113(0.037)	0.105(0.035)	0.121(0.042)	0.102(0.032)	0.108 (0.037)
	β_1	3.500(0.029)	3.499(0.024)	3.500(0.022)	3.501(0.027)	3.492(0.022)	3.497 (0.021)
	β_2	1.303(0.031)	1.298(0.022)	1.300(0.023)	1.300(0.024)	1.300 (0.024)	1.300 (0.023)

Table 4: Standard Deviations of Estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ in Example 3 (estimated standard errors in parentheses)

n		I=2		I=5		I=10		I=20	
		SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})
100	β_1	0.067	0.061(0.0088)	0.063	0.057(0.0067)	0.076	0.071(0.0073)	0.112	0.096(0.0105)
	β_2	0.067	0.062(0.0079)	0.065	0.058(0.0063)	0.086	0.071 (0.0079)	0.100	0.096(0.0112)
200	β_1	0.045	0.041(0.0037)	0.033	0.035(0.0021)	0.042	0.037(0.0024)	0.055	0.048(0.0046)
	β_2	0.041	0.041(0.0038)	0.035	0.034(0.0022)	0.039	0.037(0.0024)	0.054	0.048(0.0044)
400	β_1	0.029	0.028(0.0018)	0.024	0.023(0.0010)	0.022	0.023(0.0009)	0.027	0.025(0.0010)
	β_2	0.031	0.028(0.0018)	0.022	0.023(0.0011)	0.023	0.023(0.0009)	0.024	0.025(0.0010)

Next we test the equivalence of the two nonparametric components associated

with $Z^d = 0, 1$,

$$H_0^2 : g(Z^c, Z^d = 0) = g(Z^c, Z^d = 1),$$

$$H_1^2 : g(Z^c, Z^d = 0) \neq g(Z^c, Z^d = 1)$$

In this simulation example, $g(Z_i^c, Z_i^d = 0) = (Z_i^c)^2 + 2Z_i^c$ and $g(Z_i^c, Z_i^d = 1) = (Z_i^c)^2 + 2Z_i^c + \delta \exp(-16(Z_i^c)^2)$. To explore the relationship between effect size and power of our proposed test statistic T_n^2 , we let the value of δ change from 0 to 0.25.

To calculate T_n^2 , we first get the estimate of $\beta, \hat{\beta}$ using formula (2.5), then remove it from the model,

$$Y_i^* = g(Z_i^c, Z_i^d) + \varepsilon_i^*, \quad i = 1, \dots, n$$

where $Y_i^* = Y_i - X_i\hat{\beta}$ and $\varepsilon_i^* = \varepsilon_i + X_i\beta - X_i\hat{\beta}$. Next we use R package `locfit` to select a bandwidth h and in fact use $0.8h$ to get slightly under-smoothed estimates of $\hat{g}(Z_i^c, Z_i^d = 0)$ and $\hat{g}(Z_i^c, Z_i^d = 1)$ and their variance estimates. The test statistic T_n^2 is calculated by plugging these estimators into formula (3.16). The p -values associated with T_n^2 are calculated using the bootstrap procedure suggested in Section 3. One thousand bootstrap samples are used to approximate the null distribution of T_n^2 . This procedure is repeated 400 times to calculate the power of the test statistic under the alternative models defined by various δ values from 0 to 0.25.

The empirical distribution and bootstrapped distribution of T_n^2 under null hypothesis when $\delta = 0$ and the bootstrapped null distribution approximation under alternative hypothesis when $\delta = 0.083, 0.167, 0.25$ at sample size $n = 200$ and $I = 5$ are shown in the middle graph of Figure 4. We can see that the bootstrapped distributions under different alternative models provide fairly good approximations to the real null hypothesis distribution of our proposed test statistics. It suggests that the asymptotical null distribution of the proposed test statistics for our two population nonparametric testing problem should be a model free test statistic. In the right panel of Figure 4, the power curves of T_n^2 under various δ values and different sample

sizes are shown. The estimates of β have little impact on the power curves and such impact is only through sample size. As a two-population nonparametric test, it is not too surprising to see that the power of this test is relatively low for small sample size. But as the sample size doubles, the power function picks up quickly even for small effect size when $\delta = 0.1$.

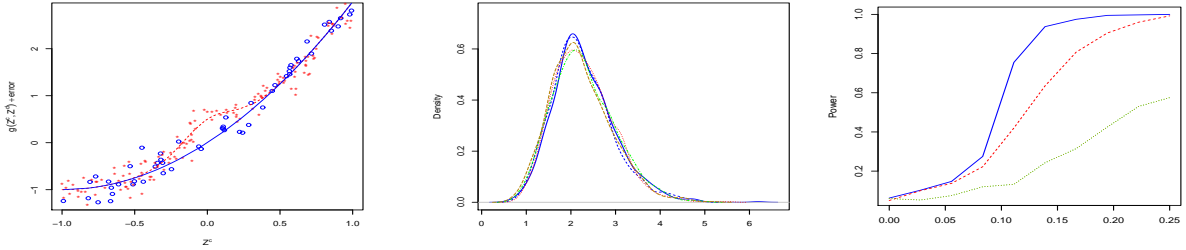


Figure 4: Example 3, Left: Scatterplot of $g(Z^c, Z^d) + \text{error}$ vs Z_2 overlaid by solid-blue line: $g(Z^c, Z^d = 0)$ and dash-red line: $g(Z^c, Z^d = 1)$. Middle: Estimated density of the empirical and bootstrap null distribution of nT_n^2 for $n = 200$ and $I = 5$: solid blue line ($\delta = 0$) is the empirical null distribution. Dash-blue line ($\delta = 0$), dot-red line ($\delta = 0.083$), dot-dash green line ($\delta = 0.167$) and long-dash dark golden red line ($\delta = 0.25$) are bootstrapped estimation of null distribution. Right: the power function evaluated at $I = 5$ and different δ values with different sample sizes $n = 100$ (dot line), $n = 200$ (dash line) and $n = 400$ (solid line).

5. Real data application: correlates of birth weight

Low birth weight is an important biological indicator since it is associated with both birth defects and infant mortality. A woman's physical condition and behavior during pregnancy can greatly affect the birth weight of the newborn. In this section, we apply our proposed methods to a classic example studying the determinants of birth weights (Hosmer and Lemeshow, 2000). This dataset is part of a larger study conducted at Bay State Medical Center in Springfield, Massachusetts. The

dataset contains variables (see below) that are believed to be associated with low birth weight in the obstetrical literatures. The goal of the analysis is to determine whether these variables are risk factors in the clinical population being served by Bay State Medical Center.

- MOTH_AGE: Mother's age (years)
- MOTH_WT: Mother's weight (pounds)
- Black: Mother's race being black ('White' is the reference group)
- Other: Mother's race being other than black or white
- SMOKE: Mother's smoking status (1=Yes, 0=No)
- PRETERM: Any history of premature labor (1=Yes, 0=No)
- HYPER: History of hypertension (1=Yes, 0=No)
- URIN_IRR: History of urinary irritation (1=Yes, 0=No)
- PHYS_VIS: Number of physician visits
- BIRTH_WT Birth weight of new born (grams)

First we analyze this data set using a linear regression model to estimate the relationship between various factors and birth weight. Shown in Table 5 (OLS-1 model), mother's race (Black vs White, Other vs White), history of pregnancy hypertension and history of urinary irritation have significantly negative impact on birth weights of newborns, while mother's weight is positively related to birth weight. Perhaps surprisingly, mother's age is not a significant predictor of baby's birth weight (p -value=0.30). To check the linearity assumption with respect to the two continuous predictors, mother's age and weight, standardized residual plot against each of them

is examined. Figure 5a shows that linearity is an adequate assumption for mother's weight and this relationship is not different between smokers and nonsmokers. But the residual diagnostics (graph not shown) indicate that the relationship between mother's age and birth weight is not linear and the relationship could potentially vary by mother's smoking status.

Then we expand the analysis to 1) a linear regression with interaction term between age and smoking (the OLS-2 model), and 2) a generalized additive model (GAM) that specifies a nonparametric term with respect to mother's age. Under the OLS-2 model, the baseline age effect is insignificant. Although the interaction term improves the model fit slightly, it is deemed insignificant (p-value=0.12). Under the GAM model, the nonparametric term of age is also tested insignificant (p-value=0.56). The conclusions about the effects of other variables on birth weights are similar compared to the OLS-1 model.

To model the nonlinear relationship between age and birth weight as well as its interaction with mother's smoking status, we further fit this data to a partially linear model with a bivariate nonparametric components, specified as,

$$\begin{aligned} \text{BirthWT} = & \beta_0 + \beta_1 \text{MOTH_WT} + \beta_2 \text{Black} + \beta_3 \text{Other} + \beta_4 \text{PRETERM} + \beta_5 \text{HYPER} \\ & + \beta_6 \text{URIN_IRR} + \beta_7 \text{PHYS_VIS} + g(\text{MOTH_AGE}, \text{SMOKE}) + \varepsilon. \end{aligned} \quad (5.1)$$

We then fit this model using the method proposed in Section 2.3. Since mother's age is recorded by a series of discrete values from 14 to 36 years, we first partition the support of $g(\text{MOTH_AGE}, \text{SMOKE})$ according to mother's smoking status, then estimate the nonparametric response curve for each group at every distinct age using available sample points (instead of using fixed cell size). The parameter estimates of the parametric components with standard errors are given in the last column of Table 5.

Table 5: Estimated effects of correlates of birth weight and their standard errors

	OLS-1	OLS-2	GAM	PL
Intercept	3026.9(308.2)	2741.9(357.0)	3044.2(309.0)	2482.0(388.2)
MOTH_WT	4.6(1.7)	4.5(1.7)	4.5(1.7)	5.6(2.0)
Black	-482.2(146.8)	-431.5(149.7)	-480.1(147.4)	-295.2(175.2)
Other	-327.5(112.6)	-302.2(113.3)	-320.1(112.9)	-203.6(132.8)
PRETERM	-179.5(133.8)	-169.8(133.4)	-166.4(134.2)	-220.0(153.5)
HYPER	-584.4(197.6)	-588.4(196.8)	-582.2(198.1)	-651.7(232.2)
URIN_IRR	-492.3(134.6)	-526.1(135.8)	-508.2(134.9)	-510.2(153.6)
PHYS_VIS	-7.0(45.4)	-0.7(45.4)	-12.2(45.5)	-14.7(52.8)
MOTH_AGE	-10.4(9.9)	1.8(12.6)	—(—)	—(—)
SMOKE	-312.5(104.5)	402.1(468.5)	-321.3(104.7)	—(—)
MOTH_AGE \times SMOKE	—(—)	-30.6(19.6)	—(—)	—(—)
R^2	0.251	0.261	0.255	0.391

Given the parametric components, we re-estimate $g(\text{MOTH_AGE}, \text{SMOKE})$ using the local polynomial regression methods via `locfit`. The fitted curves (after removing the parametric components) are shown in the right panel of Figure 5. This figure reveals that the response curves between age and birth weight are quite different for smoking and nonsmoking mothers. We can see that among non-smoking mothers, age is not particularly associated with birth weight. However, for smoking mothers, the birth weight decreases quite dramatically as mother’s age increases. The gap is as wide as over 400 grams of birth weight between nonsmoking and smoking mothers who are 30 years and older. Similar as in the simulation studies, in the local polynomial regression (`locfit`), a quadratic term is used and the

optimal bandwidth is chosen via generalized cross validation.

We also conduct the following one-sided nonparametric test to compare the two response curves between smokers and nonsmokers,

$$\begin{aligned} H_0^2 : g(\text{MOTH_AGE}, \text{Smoke}) &= g(\text{MOTH_AGE}, \text{Nonsmoke}), \quad \text{almost everywhere} \\ H_1^2 : g(\text{MOTH_AGE}, \text{Smoke}) &< g(\text{MOTH_AGE}, \text{Nonsmoke}), \quad \text{on a set with positive measure.} \end{aligned} \tag{5.2}$$

Based on (3.16), the test statistic T_n^2 for the above test is 3.36, and the bootstrap p-value is 0.029, suggesting that the response curve of age among smokers is lower than that of non-smokers. Taking this result and Figure 5b, we can see that the PL model provides a better specification for the relationship between mother's age, smoking status and birth weight.

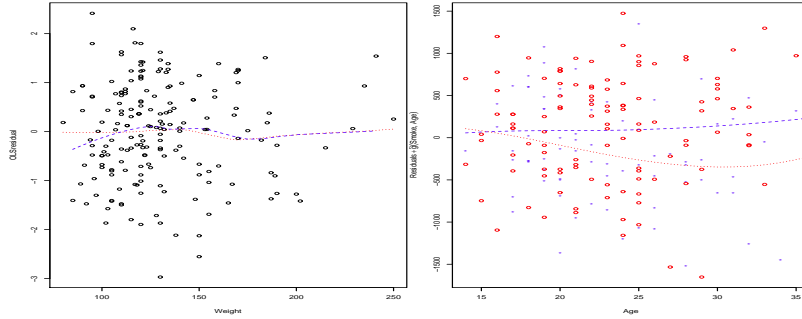


Figure 5: Correlates of birth weight: The left graph plots the residuals (OLS-2) vs mother's weight. The dotted red and dashed blue lines are the lowess fits for smoking mothers and non-smoking mothers, respectively. The right graph plots estimated regression function $g(\text{Age}, \text{Smoke})$ removing the effects of other covariates under the partial consistency PL model. The dotted red and dashed blues are the (lcofit) nonparametric estimates of response curves for smoking and non-smoking mothers, respectively.

The estimates of the parameter components of the PL model also exhibit some interesting changes compared with other models. We can see the racial gap in

birth weight narrows. Controlling other factors, on average babies born to Black mothers are 295 grams lighter than those born to White mothers. This difference is much smaller compared to the previous models. In addition, based on the test statistic defined in (3.13), the effect of "Black" now is only marginally significant (p-value=0.1) and the effect of "Other" becomes insignificant (p-value=0.147). The effect sizes and significance values of other covariates remain about the same.

6. Discussion

In this paper, based on the concept of partial consistency, we proposed a simple estimation method to partially linear regression model. The nonparametric component of the model is transformed into a set of artificially created nuisance or incidental parameters. Though these nuisance parameters cannot be estimated consistently, the parametric components of the partially linear model can be estimated consistently and almost efficiently under this configuration. As long as the sample size is reasonably large, the number of the nuisance parameters used is not too important. The estimation results have been shown to be fairly stable under various "coarseness" of the approximation. The statistical inference with respect to the parametric components via profile likelihood ratio test is also efficient. Generally speaking, the proposed simple estimation method for the partially linear regression model has two advantages that are worth noting. First, it greatly simplifies the computation burden in model estimation with little loss of efficiency. Second, it can be used to reduce the model bias by considering interaction between categorical predictor and continuous predictor, or between two continuous predictors in the nonparametric component of the model.

Though the partially linear regression model is a simple semiparametric model,

the results have offered us more insights about the “bias-efficiency” tradeoff in semiparametric model estimations: when estimating the nonparametric components, pursuing further bias reduction can increase the variance of nonparametric estimation, but it has little effect on the estimation of the parametric components of the model, and the efficient loss in the parametric part is small. Comparing to a much eased computational burden, such loss in efficiency in the parametric part can be negligible. Our study raised an interesting problem in semiparametric estimation: how to balance between the computation burden and the efficiency of the estimators while minimizing model bias. Our results can be generalized to estimate more broadly defined semiparametric models utilizing the partial consistency properties to fully exploit the information in the data.

References

- ANDREWS, D. W. K. (1994). Asymptotic for semiparametric econometric models via stochastic equicontinuity. *Econometrica*, **62** 43–72.
- CHEN, H. (1988). Convergence rates for parametric components. *Ann. Statist.*, **16** 135–146.
- CHENG, M.-Y. and WU, H.-T. (2013). Local linear regression on manifolds and its geometric interpretation. *arXiv:1201.0327* Revised for Journal of the American Statistical Association.
- ENGLE, R. F., GRANGER, C. W. J., RICE, J. and WEISS, A. (1986). Semiparametric estimates for the relation between weather and electricity sales. *J. Am. Statist. Ass.*, **81** 310–320.

- FAN, J. and HUANG, L. S. (2001). Goodness-of-fit tests for parametric regression models. *J. Am. Statist. Ass.*, **96** 640–652.
- FAN, J., PENG, H. and HUANG, T. (2005). Semilinear high-dimensional model for normalization of microarray data: A theoretical analysis and partial consistency. *J. Am. Statist. Ass.*, **100** 781–798.
- FAN, J. and ZHANG, W. (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics*, **27** 715–731.
- FAN, J. Q. and HUANG, T. (2005). Profile likelihood inferences on semiparametric varying coefficient partially linear models. *Bernoulli*, **11** 1031–1057.
- HÄRDLE, W., LIANG, H. and GAO, J. (2000). *Partially Linear Models*. Springer Verlag.
- HÄRDLE, W., MAMMEN, E. and MÜLLER, M. (1998). Testing parametric versus semiparametric modelling in generalized linear models. *J. Am. Statist. Ass.*, **93** 1461–1474.
- HECKMAN, N. E. (1986). Spline smoothing in a partly linear model. *J. R. Statist. Soc. B.*, **48** 244–248.
- HSING, T. and CARROLL, R. J. (1992). An asymptotic theory of sliced inverse regression. *Ann. Statist.*, **20** 1040–1061.
- JIANG, J., FAN, Y. and FAN, J. (2010). Estimation in additive models with highly or non-highly correlated covariates. *Ann. Statist.*, **38** 1403–1432.
- LANCASTER, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, **95** 391–413.

- LI, Q. (1996). On the root-n-consistent semiparametric estimation of partially linear models. *Econ. Lett.*, **51** 277–285.
- LIANG, H. and HÄRDLE, W. (1997). Asymptotic properties of parametric estimation in partially linear heteroscedastic models. Sonder-forschungsbereich 373 Technical report no 33, Humboldt-Universität zu Berlin.
- RACINE, J. S., HART, J. D. and LI, Q. (2006). Testing the significance of categorical predictor variables in nonparametric regression models. *Economet Rev*, **25** 523–544.
- RICE, J. (1986). Convergence rates for partially linear spline models. *Stat. Probabil. Lett.*, **4** 203–208.
- ROBINSON, P. M. (1988). Root-n consistent semiparametric regression. *Econometrica*, **56** 931–954.
- SCHICK, A. (1996). Root-n consistent estimation in partly linear regression models. *Stat. Probabil. Lett.*, **28** 353–358.
- SEVERINI, T. A. and STANISWALIS, J. G. (1994). Quasilikelihood estimation in semiparametric models. *J. Am. Statist. Ass.*, **89** 501–511.
- SPECKMAN, P. (1988). Kernel smoothing in partial linear models. *J. R. Statist. Soc. B.*, **50** 413–436.
- ZHU, L. X. and NG, K. W. (1995). Asymptotics of sliced inverse regression. *Stat Sinica* 727–736.

7. Appendix: assumptions and proofs

We need the following conditions to prove our theoretical results:

- (a). $E|\varepsilon|^4 < \infty$ and $E\|X\|^4 < \infty$.
- (b). The support of the continuous component of Z is bounded.
- (c). The functions $g(z^d, z^c)$, $E(X|Z^d = z^d, Z^c = z^c)$, the density function of Z , and their corresponding second derivatives with respect to z^c are all bounded.
- (d). Σ is nonsingular.
- (e). In presence of discrete covariate in Z , assume that for any category, the number of samples lies in this category is large enough and of order n .

For simplicity of presentation, we only discuss the case of $Z = Z^c$ and prove Theorem 1. When Z is of 2-dimension, we mainly consider that one component of Z is discrete or both components in Z are highly correlated. For the former case, according to condition (e) it can be concluded that each category has a sample size of order n . So categories do not affect the following proof which leads to the results of Corollary 1. For the latter case, assumption (2.10) implies that the following proof can be easily generalized to obtain Corollary 2. The proofs for both Corollary 1 and Corollary 2 are therefore omitted here.

Proof of Theorem 1. First, based on standard operations in least squares estimation, we can obtain the decomposition $\sqrt{n}(\hat{\beta} - \beta) = R_1 + R_2$, where

$$\begin{aligned}
 R_1 &= \left\{ \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^I \{X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i}\}^T \{X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i}\} \right\}^{-1} \\
 &\quad \times \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^J \sum_{i=1}^I \{X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i}\}^T \{g(Z_{(j-1)I+i}) - \frac{1}{I} \sum_{i=1}^I g(Z_{(j-1)I+i})\} \right\} \\
 &\equiv R_1^N / R_1^D
 \end{aligned} \tag{A.1}$$

and

$$\begin{aligned}
R_2 &= \left\{ \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^I \{X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i}\}^T \{X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i}\} \right\}^{-1} \\
&\quad \times \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^J \sum_{i=1}^I \{X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i}\}^T \{\varepsilon_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I \varepsilon_{(j-1)I+i}\} \right\} \\
&\equiv R_2^N / R_2^D
\end{aligned} \tag{A.2}$$

Hereby we will show that the term R_1 converges to zero in probability as $n \rightarrow \infty$ and the asymptotic distribution of R_2 is multivariate normal with zero mean vector and covariance matrix given in (2.11).

According to the form of R_1 , we need to first analyze the numerator R_1^N and the denominator R_1^D respectively. Let $\mathcal{F}_n = \sigma\{Z_1, Z_2, \dots, Z_n\}$ and observe that conditionally on \mathcal{F}_n , $X_{(j-1)I+i}$ are independent of each other. The following is a sketch.

We first analyze R_1^N . Denote $E(X|Z = z)$ by $m(z)$ and $X - m(Z)$ by e , then

$$\begin{aligned}
R_1^N &= \frac{1}{\sqrt{n}} \sum_{j=1}^J \sum_{i=1}^I \{m(Z_{(j-1)I+i}) - \frac{1}{I} \sum_{i=1}^I m(Z_{(j-1)I+i})\} \{g(Z_{(j-1)I+i}) - \frac{1}{I} \sum_{i=1}^I g(Z_{(j-1)I+i})\} \\
&\quad + \frac{1}{\sqrt{n}} \sum_{j=1}^J \sum_{i=1}^I \{e_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I e_{(j-1)I+i}\} \{g(Z_{(j-1)I+i}) - \frac{1}{I} \sum_{i=1}^I g(Z_{(j-1)I+i})\} \\
&= R_1^{N(1)} + R_1^{N(2)}.
\end{aligned} \tag{A.3}$$

Notice that $R_1^{N(1)}$ can be expressed using the following summations,

$$R_1^{N(1)} = \frac{1}{\sqrt{n}I^2} \sum_{j=1}^J \sum_{i=1}^I \sum_{l=1}^I \sum_{k=1}^I \{m(Z_{(j-1)I+i}) - m(Z_{(j-1)I+l})\} \{g(Z_{(j-1)I+i}) - g(Z_{(j-1)I+k})\}$$

Parallel to the proof of Hsing and Carroll (1992) and Zhu and Ng (1995), we can show that

$$\begin{aligned}
R_1^{N(1)} &\leq \frac{1}{\sqrt{n}I^2} \sqrt{\sum_{j=1}^J \sum_{i=1}^I \sum_{l=1}^I \sum_{k=1}^I \|m(Z_{(j-1)I+i}) - m(Z_{(j-1)I+l})\|^2} \\
&\quad \times \sqrt{\sum_{j=1}^J \sum_{i=1}^I \sum_{l=1}^I \sum_{k=1}^I |g(Z_{(j-1)I+i}) - g(Z_{(j-1)I+k})|^2} \\
&= O_P(n^{-1/2} I^{-2} n^\delta) = o_P(1).
\end{aligned}$$

Here δ is a arbitrarily small positive constant. Let Ω_j denote the sample set lying in the j th partition with $1 \leq j \leq J$. The last equality obtained from the fact that, under condition (c), $m(\cdot)$ and $g(\cdot)$ have a total variation of order δ ,

$$\begin{aligned}
\lim_{n \rightarrow \infty} \frac{1}{n^\delta} \sup_{\{\Omega_j, 1 \leq j \leq J\}} \sum_{i=1}^{I-1} \|m(Z_{(j-1)I+i}) - m(Z_{(j-1)I+(i+1)})\| &= 0, \\
\lim_{n \rightarrow \infty} \frac{1}{n^\delta} \sup_{\{\Omega_j, 1 \leq j \leq J\}} \sum_{i=1}^{I-1} |g(Z_{(j-1)I+i}) - g(Z_{(j-1)I+(i+1)})| &= 0.
\end{aligned}$$

Next we consider $R_1^{N(2)}$. Let $\bar{e}_{(n)}$ and \bar{e}_1 be the largest and smallest of the corresponding e_i 's, respectively. It is clear that

$$\begin{aligned}
R_1^{N(2)} &\leq \frac{\bar{e}_{(n)} - \bar{e}_1}{\sqrt{n}I} \sum_{j=1}^J \sum_{i=1}^I \sum_{l=1}^I |g(Z_{(j-1)I+i}) - g(Z_{(j-1)I+l})| \\
&= 2 \frac{\bar{e}_{(n)} - \bar{e}_1}{\sqrt{n}I} \sum_{j=1}^J \sum_{1 \leq i < l \leq I} |g(Z_{(j-1)I+i}) - g(Z_{(j-1)I+l})|
\end{aligned}$$

The above argument leads to that

$$\begin{aligned}
R_1^{N(2)} &\leq 2 \frac{\bar{e}_{(n)} - \bar{e}_1}{\sqrt{n}I} \sum_{i=1}^I \sum_{l=1}^I \sum_{j=1}^{n-1} |g(Z_{(j+1)}) - g(Z_{(j)})| \\
&\leq 2I \frac{\bar{e}_{(n)} - \bar{e}_1}{\sqrt{n}} \sum_{j=1}^{n-1} |g(Z_{(j+1)}) - g(Z_{(j)})|.
\end{aligned}$$

Applying Lemma A.1 of Hsing and Carroll (1992), we obtain

$$n^{-1/4} |\bar{e}_{(n)} - \bar{e}_1| \xrightarrow{P} 0.$$

Note the fact that total variation of $g(\cdot)$ is of order n^δ , we have $R_1^{N(2)} = o_P(1)$.

Combining the results about $R_1^{N(1)}$ and $R_1^{N(2)}$, the proof for R_1^N is completed.

Next consider R_1^D and R_2^D . Since $R_1^D = R_2^D$, we only need to show the case of R_1^D . The expectation of R_1^D is calculated as follows.

$$\begin{aligned} \mathbb{E}(R_1^D) &= \mathbb{E}(XX^\top) - \frac{1}{nI} \sum_{j=1}^J \sum_{i=1}^I \sum_{l=1}^I \mathbb{E}\{X_{(j-1)I+i}X_{(j-1)I+l}\} \\ &= \mathbb{E}(XX^\top) - \frac{1}{nI} \sum_{j=1}^J \sum_{i=1}^I \mathbb{E}\{X_{(j-1)I+i}X_{(j-1)I+i}\} - \frac{1}{nI} \sum_{j=1}^J \sum_{i \neq l}^I \mathbb{E}\{X_{(j-1)I+i}X_{(j-1)I+l}\} \\ &= (1 - \frac{1}{I}) \mathbb{E}(XX^\top) - \frac{1}{nI} \sum_{j=1}^J \sum_{i \neq l}^I \mathbb{E}\left[\mathbb{E}\{X_{(j-1)I+i}X_{(j-1)I+l}|\mathcal{F}_n\}\right] \end{aligned}$$

Under the assumption that conditionally on \mathcal{F}_n , $X_{(j-1)I+i}$ are independent of each other, we can obtain that $\mathbb{E}\{X_{(j-1)I+i}X_{(j-1)I+l}|\mathcal{F}_n\} = m(Z_{(j-1)I+i})m(Z_{(j-1)I+l})$.

This, together with the above analysis, gives

$$\begin{aligned} \mathbb{E}(R_1^D) &= (1 - \frac{1}{I}) \mathbb{E}(XX^\top) - \frac{I-1}{nI} \sum_{j=1}^J \sum_{i=l}^I \mathbb{E}\left[m(Z_{(j-1)I+i})m(Z_{(j-1)I+i})\right] \\ &\quad - \frac{1}{nI} \sum_{j=1}^J \sum_{i \neq l}^I \mathbb{E}\left[m(Z_{(j-1)I+i})\{m(Z_{(j-1)I+l}) - m(Z_{(j-1)I+i})\}\right] \\ &= (1 - \frac{1}{I}) \mathbb{E}(XX^\top) - \frac{I-1}{nI} \sum_{j=1}^J \sum_{i=l}^I \mathbb{E}\left[m(Z_{(j-1)I+i})m(Z_{(j-1)I+i})\right] + o(1) \\ &= (1 - \frac{1}{I}) \mathbb{E}\left[\{X - \mathbb{E}(X|Z)\}\{X - \mathbb{E}(X|Z)\}^\top\right] + o(1). \end{aligned}$$

The term of order $o(1)$ is obtained following a similar argument of Theorem 2.3 of Hsing and Carroll (1992). This completes the proof for R_1 .

We now deal with the term R_2 . Observe that given $\{(X_i, Z_i), i = 1, \dots, n\}$, each term of $\{\varepsilon_{(j-1)I+i} - \frac{1}{J} \sum_{j=1}^J \varepsilon_{(j-1)I+i}\}$ has mean zero and is independent of each other. Thus R_2 is asymptotically normal with mean zero. We will show that the limiting variance of R_2 is equal to the covariance matrix given in (2.11). That is,

$$\begin{aligned} \text{Var}(R_2|\{X_i, Z_i\}) &= (R_2^D)^{-1} \text{Var}(R_2^N|\{X_i, Z_i\})(R_2^D)^{-1} \\ &= \{\mathbb{E}(R_2^D)\}^{-1} \mathbb{E}\{\text{Var}(R_2^N|\{X_i, Z_i\})\} \{\mathbb{E}(R_2^D)\}^{-1} + o_P(1) \end{aligned}$$

and

$$\begin{aligned}
\text{Var}(R_2^N | \{X_i, Z_i\}) &= \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^I \{X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i}\} \{X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i}\}^\top \\
&\quad \times \mathbb{E} \left[\left\{ \varepsilon_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I \varepsilon_{(j-1)I+i} \right\}^2 \middle| \{X_i, Z_i\} \right] \\
&= \frac{\sigma^2}{n} \sum_{j=1}^J \sum_{i=1}^I \{X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i}\} \{X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i}\}^\top \\
&\xrightarrow{P} \sigma^2 \left(1 - \frac{1}{I}\right) \mathbb{E} \left[\{X - \mathbb{E}(X|Z)\} \{X - \mathbb{E}(X|Z)\}^\top \right].
\end{aligned}$$

Combining the last two equations, we complete the proof of Theorem 1. \square

Proof of Theorem 2. First we show that $\text{RSS}_1 = \sigma^2\{1 + o_P(1)\}$. By (2.5),

$$\begin{aligned}
\text{RSS}_1 &= \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^I \{Y_{(j-1)I+i} - \hat{\alpha}_{j1} - X_{(j-1)I+i} \hat{\beta}_1\}^2 \\
&= \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^I \left[\{X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i}\} (\beta - \hat{\beta}_1) \right. \\
&\quad \left. + \{g(Z_{(j-1)I+i}) - \frac{1}{I} \sum_{i=1}^I g(Z_{(j-1)I+i})\} \right. \\
&\quad \left. + \{\varepsilon_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I \varepsilon_{(j-1)I+i}\} \right]^2 \\
&= I_1 + I_2 + I_3 + I_4 + I_5 + I_6,
\end{aligned}$$

where

$$\begin{aligned}
I_1 &= \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^I \left\{ \varepsilon_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I \varepsilon_{(j-1)I+i} \right\}^2 \\
I_2 &= \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^I \left[\left\{ X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i} \right\} (\beta - \hat{\beta}_1) \right]^2 \\
I_3 &= \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^I \left\{ g(Z_{(j-1)I+i}) - \frac{1}{I} \sum_{i=1}^I g(Z_{(j-1)I+i}) \right\}^2 \\
I_4 &= \frac{2}{n} \sum_{j=1}^J \sum_{i=1}^I \left\{ \varepsilon_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I \varepsilon_{(j-1)I+i} \right\} \left\{ g(Z_{(j-1)I+i}) - \frac{1}{I} \sum_{i=1}^I g(Z_{(j-1)I+i}) \right\} \\
I_5 &= \frac{2}{n} \sum_{j=1}^J \sum_{i=1}^I \left\{ X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i} \right\} (\beta - \hat{\beta}_1) \left\{ g(Z_{(j-1)I+i}) - \frac{1}{I} \sum_{i=1}^I g(Z_{(j-1)I+i}) \right\} \\
I_6 &= \frac{2}{n} \sum_{j=1}^J \sum_{i=1}^I \left\{ X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i} \right\} (\beta - \hat{\beta}_1) \left\{ \varepsilon_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I \varepsilon_{(j-1)I+i} \right\}
\end{aligned}$$

Using the same arguments when analyzing R_1 and R_2 , it can be shown that

$$\begin{aligned}
I_1 &= \frac{I-1}{I} \sigma^2 \{1 + o_P(1)\}, & I_2 &= O_P(n^{-1}), & I_3 &= o_P(n^{-1/2}), \\
I_4 &= o_P(n^{-1/4}), & I_5 &= o_P(n^{-3/4}), & I_6 &= o_P(n^{-1/2}).
\end{aligned}$$

Similarly, RSS_0 can be decomposed as

$$\begin{aligned}
\text{RSS}_0 &= \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^I \{Y_{(j-1)I+i} - \hat{\alpha}_{j0} - X_{(j-1)I+i} \hat{\beta}_0\}^2 \\
&= \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^I \left[\{Y_{(j-1)I+i} - \hat{\alpha}_{j1} - X_{(j-1)I+i} \hat{\beta}_1\} \right. \\
&\quad \left. + \{X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i}\} (\hat{\beta}_1 - \hat{\beta}_0) \right]^2 \\
&= \text{RSS}_1 + J_1 + J_2,
\end{aligned}$$

with

$$\begin{aligned}
J_1 &= \frac{1}{n} \sum_{j=1}^J \sum_{i=1}^I \left[\left\{ X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i} \right\} (\hat{\beta}_1 - \hat{\beta}_0) \right]^2 \\
J_2 &= \frac{2}{n} \sum_{j=1}^J \sum_{i=1}^I \{Y_{(j-1)I+i} - \hat{\alpha}_{j1} - X_{(j-1)I+i} \hat{\beta}_1\} \left\{ X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i} \right\} (\hat{\beta}_1 - \hat{\beta}_0).
\end{aligned}$$

From the proof of Theorem 1, it holds that $\frac{1}{n} \sum_{j=1}^J \sum_{i=1}^I \{X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i}\} \{X_{(j-1)I+i} - \frac{1}{I} \sum_{i=1}^I X_{(j-1)I+i}\}^\top \xrightarrow{P} \frac{I-1}{I} \Sigma$. Furthermore, the estimators for β under the null and alternative hypotheses then have the following relation

$$\hat{\beta}_0 = \hat{\beta}_1 - \Sigma^{-1} A^\top \{A \Sigma^{-1} A^\top\}^{-1} A \hat{\beta}_1 + o_P(\hat{\beta}_1).$$

J_1 can then be written as

$$J_1 = \frac{I-1}{I} \hat{\beta}_1^\top A^\top \{A \Sigma^{-1} A^\top\}^{-1} A \hat{\beta}_1 + o_P(\hat{\beta}_1).$$

This, together with the asymptotic normality of $\hat{\beta}_1$ in Theorem 1 implies that under the null hypothesis $A \hat{\beta}_1 \xrightarrow{\mathcal{L}} N(0, \sigma^2 \frac{I-1}{I} A \Sigma^{-1} A^\top)$, we have $nJ_1 \xrightarrow{\mathcal{L}} \sigma^2 \chi_k^2$. By some calculation, it can be shown that $J_2 = 0$. Thus,

$$n(\text{RSS}_0 - \text{RSS}_1) \xrightarrow{\mathcal{L}} \sigma^2 \chi_k^2.$$

Then by Slutsky theorem,

$$nT_n^1 = n \frac{\text{RSS}_0 - \text{RSS}_1}{\text{RSS}_1} \xrightarrow{\mathcal{L}} \frac{I}{I-1} \chi_k^2. \quad \square$$